H-009

段階的学習による物体検知向け半自動アノテーション
Semi-automatic annotation for object detection by stepwise learning

井下 哲夫† 石井 遊哉† 中野 学† 高橋 勝彦†
Tetsuo Inoshita Asuka Ishii Gaku Nakano Katsuhiko Takahashi

1. はじめに

近年，深層学習の発展により画像認識精度は飛躍的に向上している．一般的に，精度を向上させるためには，学習データと呼ばれる画像と正解ラベルの組を，大量に準備しておくことが必要になる．しかしながら，学習データを大量に準備するためには，収集した画像に対して正解ラベルを付与するアノテーション作業に膨大な工数が発生する．アノテーション作業を効率化する技術は様々な手法が提案されており，完全手動，半自動の 2 つに分類される．

1) 完全手動：完全手動のアノテーションでは矩形描画ツールを用いて正解ラベル（外接矩形やポリゴン）を付ける方法が一般的である．Web ベースや Matlab Tool Box 形式で提供されている LabelMe[1]や，物体検知アルゴリズムで著名な SSD[2]や YOLO[3]向けのアノテーションフォーマットを出力する labelImg[4]が公開されている．

2) 半自動：機械学習にとって重要なデータのみ人がアノテーションを行う方法で，重要なデータかどうかの判断には一般的に能動学習（Active Learning[5]）が用いられる．また半教師あり学習の一つである self-training 法[6]や，ラベルなしデータに付与される仮ラベルを自動評価することで，衛星画像認識に有効な学習データを収集する方法が提案されている[7].

しかしながら，いずれの方法も入手できるデータや課題設定によって効果が変わるため，解くべき問題に対応した工夫が必要になる．本研究では，映像監視における物体検知問題を対象としている．そのため，入手できるデータは監視カメラ等の連続した映像データであり，前後フレームの関係性が強い．例えば，あるフレームで登場する人や物はその前後フレームでも登場している確率が高い．従って，代表的なフレームを選択し学習することができれば，残りのフレームも精度良く self-training することができる．

本論文では，1)2)の長所を取り入れたアノテーション方法を提案する．具体的には，分割したデータセットから，データセットを代表する数枚のデータを選択する．その少数のデータのみ手動で正解ラベルを付与，学習し，残りのデータを推論させる．推論結果は人手で確認，修正を行う．この工程を分割数ごとに段階的に進めていくことで，大量の映像データに対するアノテーション作業を効率化できる．実映像を用いた実験により，提案手法によって品質（精度）は維持したまま，1/3 程度の作業時間削減を確認した．

2. 提案手法

物体検知タスクにおける学習データとは，一般的には動画から出力した画像とその中に含まれる物体の位置座標（外接矩形）を記した正解ラベルである．本論文ではこれらの正解ラベルを生成することを目的とし，提案する物体検知向けアノテーション手法の処理の流れを図 1 に示す．以下図 1 を参照しながら説明する．

2.1 データ分割・学習用シード画像の選択

監視カメラから取得した学習用のデータセットは動画であることが多い．提案手法では動画の特性であるフレーム間の連続性を活かすため，データセットを等分割ではなく段階的に増えるように「分割」する(図 1 では 2 分割)．次に「選択」では，学習時のシードとなる n 枚のデータを選択する．選択方式としてランダムサンプリングや一定間隔でのサンプリングが考えられるが，ここでは，RGB 値による k-means clustering を用いてセントロイドとなる画像を n 枚選択する．その後，選択した n 枚に対して正解ラベルを人手で付け，初期の学習用シード画像として利用する．

2.2 学習・推論・修正・多段階処理

n 枚の正解ラベルを用いて学習を行い，残りのデータセットに対して推論を行う．これを複数回行う．推論結果に対する修正戦略は，信頼度に基づく least confidence[5]を採用した．なお推論結果は，物体の外接矩形座標値と予測した物体名から構成されている．提案手法では，分割したデータセットに対してのみ推論を行うので，過学習気味に学習パラメータを調整する必要がある．修正したラベルは正解ラベルとして，次段階の学習データとなり，残りの未アノテーションデータを推論する際に使用する．この処理を分割したデータセット数分繰り返し，最終的に正解ラベルが付いた学習データが生成される．

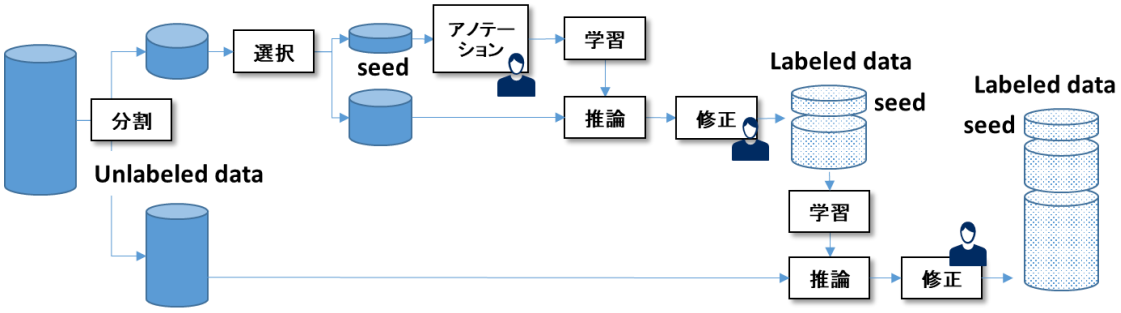

図 1　提案手法の処理の流れ

† NEC 中央研究所　NEC　Central Research Labs.


data1(1043 枚, 5169 個)


data2(1166 枚, 11559 個)

図 2 実験画像枚数とアノテーション数

3. 実験

3.1 実験条件

「完全手動」と「提案手法」によりアノテーションを実施した場合について，作業時間と精度について比較する．実験で用いるデータは，独自の持ち物検知用データセット 2 種類（data1, data2）で，アノテーション対象の物体数は人，カバン，杖等の持ち物 9 種類である（図 2）．

評価尺度は，アノテーションに要した処理時間とそのアノテーション結果を用いて，各データセットに対して 2-fold cross validation の物体検知精度（mAP: mean average precision）を用い，9 種類の平均 mAP を評価する．なお，ラベル付けの偏りを考慮するために，完全手動，提案手法の両方で作成したデータを評価データに用いる．

また，物体検知アルゴリズムは Retinanet[6]を用い，アノテーション作業は一般人（技術専門外）2 名で行った．

3.2 実験手順

完全手動での実験手順は，弊社で開発した GUI ツールを用いて人手によるアノテーション作業を行った．

提案手法での実験手順を図 1 に基づいて説明する．まず，data1（1043 枚）では，2-fold cross validation 用にデータを 2 つに分割した後，1 つのデータセット（520 枚）から 50 枚，残り 470 枚に「分割」する．50 枚のデータセットから 10 枚を「選択」し人手によるアノテーションを行う．その後，10 枚を「学習」し，40 枚に対して「推論」を行う．推論結果を人手で「修正」し，1 段階目の学習データ 50 枚を作成する．次に 2 段階目として 50 枚のラベルを更に「学習」し，残り 470 枚に対して「推論」を行う．推論結果を人手で「修正」し，学習データを作成する．最後に，作成した学習データを用いて cross validation を行い，精度評価を行う．

4. 結果

表 1 に評価結果を示す．data1 の場合，完全手動と提案手法とでは，平均 mAP は変わらず，作業時間は 37 時間から 11 時間 27 分になり，約 1/3 程度削減できていることが分かる．また，data2 も同様に，平均 mAP は変わらず，作業時間は 72 時間 20 分から 25 時間 33 分になり，約 1/3 削減できている．

提案手法によるアノテーションは，物体検知アルゴリズムによる検知結果を用いるため，一部の「人」ラベルについては完全手動と比べると位置ずれが発生している．data1,2 での位置ずれ量を IoU (Intersection of Union)値で計測すると平均 0.86 であり，特に大きな位置ずれ例を図 2 に示す．しかしながら検知精度は変わらない．これは，学習時に正解ラベルの物体位置情報を微小に摂動（data augmentation）させているため，位置ずれの影響が吸収されていると考えられる．このことから，人手によるアノテーションでも，厳密に物体の外接矩形を設定する必要がないことが分かる．

表 1　評価結果

評価データ		完全手動	提案手法
data1	mAP	0.59	**0.60**
	作業時間	37h	**11h 27m**
data2	mAP	0.61	**0.63**
	作業時間	72h 20m	**25h 33m**

図 3 完全手動と提案手法とのラベル位置ずれ例
（左：完全手動，右：提案手法）

また，実験に使用した画像群は，映像の一部であるため時系列的な関係性が強い．つまり一定枚数の画像の中に同一人物や，同じような物体が映っていることから，代表的な 10 枚の中に残りデータセットの情報が網羅できている可能性がある．従って，提案手法では時系列映像には効果があるが，様々な場所で撮影された静止画群がデータセットの場合には効果が小さくなると考えられる．

5. おわりに

本論文では，監視映像中の物体検知タスクを想定し，最初にデータセットの 10 枚を人手でアノテーションした後，残りのデータセットに対して，段階的に学習・推論を行うことで，全て人手でアノテーションする場合に比べて，品質は維持しながらも作業時間が 1/3 に削減できることを示した．今後は 10 枚の選択方法の改良やラベルなしデータを活用することで，更なる効率化を目指していく．

参考文献

[1] B. Russell, A. Torralba, K. Murphy, W. T. Freeman, “LabelMe: a database and web-based tool for image annotation”, International Journal of Computer Vision, pages 157-173, Volume 77, Numbers 1-3, May, 2008
[2] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, and A. C. Berg, “SSD: Single shot multibox detector”, In European Conference on Computer Vision, pages 21–37. Springer, 2016.
[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi,"You only look once: Unified, real-time object detection", arXiv preprint arXiv:1506.02640, 2015.
[4] Tzutalin, https://github.com/tzutalin/labelImg. git code, 2015
[5] Burr Settles, “Active Learning Literature Survey”, Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.
[6] Zhu, X.: Semi-Supervised Learning Literature Survey, “Technical Report 1530”, Computer Science, University of Wisconsin-Madison, 2005
[7] 上原, 野里, 村川, 坂無, ”人と協調する半自己学習に基づく衛星画像上の地物検出”, CVIM-216, No.9, 2019.
[8] T-Y Lin, P Goyal, R Girshick, K He, P Dollár. Focal Loss for Dense Object Detection. IEEE International Conference on Computer Vision (ICCV), 2017.